УДК 004.5:81'42

Я. В. Поточняк

# ОБРАБОТКА ТЕКСТОВ И СООБЩЕНИЙ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ В ИНФОРМАЦИОННЫХ СИСТЕМАХ

*В статье рассматриваются вопросы использования ограниченного естественного языка для общения конечного пользователя с информационной системой, а также обработки текстовых документов для автоматизированного обмена информацией с базами данных.*

**Ключевые слова:** *термин, устойчивое словосочетание, анализ текста, словарь предметной области.*

## 1. Введение

Исследования, представленные в данной работе, относятся к области искусственного интеллекта и автоматизированного документооборота. В крупных современных организациях циркулирует огромное количество документов. Наряду с хранением электронных вариантов этих документов часто нужно извлекать из них отдельные данные для размещения в базе данных (БД). Существует и обратная задача — извлечения из БД нужной информации неквалифицированным пользователем. В настоящее время эти процессы практически не автоматизированы. Поэтому исследования в данной области следует считать актуальными.

## 2. Постановка проблемы

Каждая организация работает в определенной предметной области, поэтому первой решаемой проблемой является автоматизация процесса построения словаря конкретной предметной области. Затем следует установить связь между содержимым БД и словарем и, наконец, обеспечить анализ текстовых документов или запросов конечного пользователя на ограниченном естественном языке для обмена информацией с БД.

## 3. Основная часть

**3.1. Анализ литературных источников по теме исследования.** В работе [1] предложено использовать объектный словарь данных для отображения содержимого реляционной БД с целью формирования запросов к БД на языке OSQL (объектный SQL), позволяет ускорить процесс разработки прикладного программного обеспечения в ИС.

В работе [2] рассмотрена возможность выполнения семантического анализа текста в автоматизированной обучающей системе на базе программного продукта ДИАЛИНГ и формирование семантического «образа» текста с помощью сети фреймов. Предложенные решения позволяют оценить ответы обучаемого на естественном языке.

В работе [3] предложено организовать двухуровневый словарь сущностей, где первый (внешний) уровень представляет сущности предметной области, их атрибуты и связи между сущностями — ассоциации. Второй уровень предусматривает ссылки на физическое представление данных в базе данных информационной системы. Такое представление данных позволяет производить интеграцию распределенных систем на уровне словаря.

В работе [4] предложен метод формирования шаблонов запросов конечного пользователя к РБД, что позволяет последнему решать многие задачи без привлечения прикладного программиста. Метод базируется на использовании объектного словаря БД, который ограничивает пользователя в выборе получаемой информации.

В работе [5] разработана поэтапная технология формирования SQL-запроса к РБД на основании обращения пользователя к системе на ограниченном естественном языке и его диалога с системой. Технология основана на использовании объектного словаря БД.

В работе [6] предложено провести лексический, морфологический, синтаксический и семантический анализ документов некоторой организации для создания словаря соответствующей предметной области. Разработана технология исключения из рассмотрения слов, которые не могут являться терминами. Предложены правила формирования многословных терминов (именных групп).

В работе [7] на основании анализа структур, которые используются для интеллектуальной обработки данных, была разработана модель для извлечения и сохранения фактов на естественном языке, что является основой для проектирования ИС с естественно-языковым интерфейсом. Предложен метод сравнения моделей двух текстов для решения задач поиска и классификации объектов.

В работе [8] получил дальнейшее развитие метод построения словаря предметной области [6].

Предложена предварительная классификация текстов, введены весовые коэффициенты важности текста, решается вопрос о достаточном количестве обрабатываемой информации.

**3.2. Результаты исследований.** Исследования применимости словаря предметной области (СПП), построенного в соответствии с [6, 8] показал, что в реальной ИС для работы с документами необходимо расширить его возможности. Соответственно были разработаны алгоритмы для выделения из исходных текстов и занесения в словарь аббревиатур, некоторых сокращений, типичных для данной предметной области названий (организаций, министерств и т. д.).

Другое направление исследований направлено на интеграцию СПП с БД, используемой в ИС. В работах [1, 3, 5] задача общения конечного пользователя с БД на ограниченном естественном языке решалась с помощью объектного словаря БД, который строился на основании анализа структуры БД и последующего подбора терминов, соответствующих хранимой в БД информации. Автоматизированный анализ структуры БД, позволил установить в СПП ссылки на соответствующие элементы БД. Таким образом содержимое СПП, построенного на основании анализа множества текстовых документов конкретной организации, перекрывает возможности объектного словаря БД.

Предложенный СПП является основой для классификации и распознавания документов, оценки возможности сохранения в БД или извлечения из БД данных, связанных с конкретным документом, создания шаблонов документов, построения интерфейса пользователя на ограниченном естественном языке.

Показано, что сравнение СПП для различных ИС может служить мерой и основой возможной интеграции этих систем.

### Литература


1. Кунгурцев А. Б. Использование словаря данных в информационных системах с логической формой представления данных [Текст] / **А. Б. Кунгурцев**, А. А. Завалин // Труды ОНПУ. — 2002. — Вып. № 2(18). — С. 121—126.
2. Кунгурцев А. Б. Средства представления и анализа ответов обучаемого на естественном языке [Текст] / **А. Б. Кунгурцев**, Е. А. Пиринова, Н. А. Новикова // Труды Украинской научно-методической конференции «Нові інформаційні технології навчання в навчальних закладах України». — Одесса : ОДМУ. — 2003. — № 9, частина 2. — С. 236—242.
3. Кунгурцев А. Б. Формирование представления даннях распределенных информационных систем в терминах предметной области [Текст] / **А. Б. Кунгурцев** // Нові технології. — 2003. — № 2(3). — С. 74—77.
4. Кунгурцев О. Б. Формування шаблонів запитів для реляційних баз даних з використанням об'єктного словника [Текст] / **О. Б. Кунгурцев**, І. В. Барикіна, О. А. Завалін // Наукові праці ОНАХТ. — Одеса. — 2004. — № 27. — С. 233—236.
5. Кунгурцев А. Б. Методика формирования запросов к реляционной базе данных конечным пользователем [Текст] / **А. Б. Кунгурцев** // Искусственный интеллект. — 2004. — № 1 — С. 60—65.
6. Кунгурцев А. Б. Формирование словаря предметной области [Текст] / **А. Б. Кунгурцев**, И. В. Тыхан // Искусственный интеллект. — 2006. — № 1. — С. 144—151.
7. Кунгурцев О. Б. Застосування мереж фреймів для побудови моделі вилучення фактів з текстів природною мовою [Текст] / **О. Б. Кунгурцев**, С. М. Бородавкін // Штучний інтелект. — 2009. — № 4. — С. 202—207.
8. Кунгурцев А. Б. Метод построения словарей предметных областей для извлечения фактов из текстов на естественном языке [Текст] / **А. Б. Кунгурцев**, С. Н. Бородавкин, А. П. Голуб // Восточно-Европейский журнал передовых технологий. — 2010. — № 1/4(43). — С. 32—36.


## ОБРОБКА ТЕКСТІВ І ПОВІДОМЛЕНЬ НА ПРИРОДНІЙ МОВІ В ІНФОРМАЦІЙНИХ СИСТЕМАХ

**Я. В. Поточняк**

У статті розглядаються питання використання обмеженої природної мови для спілкування кінцевого користувача з інформаційною системою, а також обробки текстових документів для автоматизованого обміну інформацією з базами даних.

**Ключові слова:** термін, стійке словосполучення, аналіз тексту, словник предметної області.

***Яна Володимирівна Поточняк**, студентка шостого курсу (магістратура), кафедра системного програмного забезпечення Одеського національного політехнічного університету, **тел.: 0661322363, e-mail: yana_onpu@mail.ru**.*

## THE PROCESSING TEXT AND MESSAGES ON THE NATURAL LANGUAGES IN INFORMATION SYSTEMS

**I. Potochniak**

The article examine questions of using limited natural language. It needed for communications of end-user with an information system. And also for processing text documents for automated information exchanging with databases.

**Keywords:** term, sustainable phrase, text analysis, dictionary domain.

***Iana Potochniak**, a student of the sixth course (Master), Department of System Software Odessa National Polytechnic University, **tel.: 0661322363, e-mail: yana_onpu@mail.ru**.*